Physical Ability-Task Performance Models: Assessing the Risk of Omitted Variable Bias

Ross R. Vickers, Jr.
James A. Hodgdon
Marcie B. Beckett

Naval Health Research Center

Document No. 09-04

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. Approved for public release; distribution is unlimited.

This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Naval Health Research Center
140 Sylvester Rd.
San Diego, California 92106-3521

Physical Ability-Task Performance Models:
Assessing the Risk of Omitted Variable Bias

Ross R. Vickers, Jr.
James A. Hodgdon

Warfighter Performance Department
Naval Health Research Center
140 Sylvester Road
San Diego, CA 92106-3521

and

Marcie B. Beckett
San Diego, CA

Report No. 09-04 was supported by the Navy Medical Research and Development Command under research Work Unit No. 63706N M0096.02-1050 and by the Office of Naval Research, Arlington, VA, under Work Unit No. 60704. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution is unlimited. Human subjects participated in this study after giving their free and informed consent. This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Abstract

The physical capacities of job incumbents limit performance on occupational physical tasks. While muscle strength is logically an important performance-relevant physical ability, omitted variable bias may cause its importance to be overstated. This bias occurs when a causal variable in a model correlates with other causal variables that are omitted from the model. The impact of omitted variable bias on the strength-performance association was evaluated in a study of simulated job performance in men and women. The study measured four major abilities, Static Strength (SS), Dynamic Strength (DS), Anaerobic Power (AP), and Aerobic Capacity (AC). Performance measures were simulated lifting and carrying tasks. Analysis showed moderate to strong relationships among the ability measures. All four ability measures were significantly related to lifting and to carrying performance. However, construction of a series of alternative predictive models led to adoption of a final model, with SS and AC as the only predictors. The absence of AP and DS from the model indicates that omitted variable bias can be expected whenever these ability factors are studied in isolation from SS and AC. The practical implication is that physical training can be mistakenly focused on abilities that have no impact on job performance.

The physical capacities of job incumbents limit their performance on physically demanding occupational tasks. This person-task interplay is important for selection practices and job design. Muscle strength is a critical physical ability. The correlation between strength test performance and physical task performance regularly exceeds *r* = .85 (Arnold, Rauschenberger, Soubel, & Guion, 1982; Hogan, 1991a) and can exceed *r* = .90 (Vickers, 1995, 1996). These relationships are strong enough to suggest that muscle strength is the only ability that must be considered for personnel selection and job design.

Omitted variable bias may have inflated the apparent muscle strength-task performance association in previous studies. This bias occurs when a causal variable in a model is correlated with other causal variables that are missing from the model (James, Mulaik, & Brett, 1982). In this case, part of the effects of the omitted variables will be attributed to the included variable. Bias is a concern in the present context because muscle strength is correlated with other physical abilities (e.g., Myers, Gebhardt, Crump, & Fleishman, 1993). Bias, therefore, will occur if these other abilities affect task performance and are omitted from the model.

Vickers (2003a) demonstrated the potential for omitted variable bias in the estimates of strength effects in a reanalysis of Arnold et al.'s (1982) steelworker data. Arnold et al. (1982) included variables representing two strength dimensions in their study. One dimension, static strength (SS), corresponds to the usual concept of general muscle strength (cf., Vickers, 2003b), which is generally defined as the maximum force that a muscle can generate (Kroemer, Kroemer, & Kroemer-Elbert, 1990). The other dimension, dynamic strength (DS), is closer to the concept of muscle endurance, which corresponds to the continuous or repetitive duration of submaximal exertions. These strength dimensions were positively correlated (*r* = .76), so one essential condition for the occurrence of omitted variable bias was satisfied. The reanalysis also indicated that DS was correlated with performance (*r* = .76), so the second essential condition for the occurrence of omitted variable bias was satisfied. In the reanalysis, two structural models were fitted to Arnold et al.'s (1982) data. In the first model, SS was the only predictor of performance. In the second model, DS was added to the predictive model. Adding DS to the model reduced the standardized regression slope for SS by ~20% (i.e., from β = .86 to β = .69). This outcome indicates that omitted variable bias can be substantial when evaluating the relationship between physical abilities and physical task performance.

The reanalysis of Arnold et al.'s (1982) data also illustrates a second type of bias. Confirmation bias occurs when a less-than-optimal model is accepted as satisfactory because it is not compared with alternative models (cf., MacCallum & Austin, 2000). Thus, this second type of bias involves the search for alternative models and decisions about the adequacy of any given model. Based on this example, confirmation bias can occur when physical abilities are investigated one at a time.

This study extended the investigation of the risk of bias when modeling the physical ability-performance domain. Both sides of the ability-performance equation were extended. Two physical ability dimensions, anaerobic power (AP) and aerobic capacity (AC), were added to the ability profile. On the criterion side, brief (i.e., <1 min) lifting tasks and moderate duration carrying tasks were treated as separate performance dimensions. Earlier work involved either brief tasks (Vickers, 1995, 1996) or moderate duration (5 min to 15 min) tasks (Vickers, 2003a), but not both. Models that considered all four ability dimensions as predictors of both performance dimensions were compared to better evaluate the complexity of the ability-performance interface for physical tasks.

Methods

Sample

Participants were active-duty naval personnel (64 men, 38 women) between ages 20 and 35 years. Each participant passed a screening test in which the individual stood upright and pulled on the handles of a small metal box held at knuckle height. The box was attached to a dynamometer (model TCG-500, John Chatillon & Sons, New York, NY) that measured the maximal force exerted. Individuals were permitted to participate only if they could generate at least 76-kg lifting force. This force was the minimum required to ensure safe performance of the job task simulations. The minimum value was determined from Monod’s (1985) equation for the strength requirements for intermittent static work (cf., Beckett & Hodgdon, 1987, for additional details).

The data analyzed for this report were obtained from 93 of the 102 participants admitted to the study. The other 9 participants had missing data for one or more of the variables used in the present analyses.

Ability Measures

The ability measures were a subset of a larger battery of measures collected by Beckett and Hodgdon (1987). The subset was chosen to cover the strength and endurance domains and to be comparable to the measures used in prior models of data from Robertson and Trent (1985) and Arnold et al. (1982). Measures were:

> *Incremental Lift Machine* (ILM). The ILM consisted of an adjustable weight stack, a lift bar for moving the stack, and two upright tracks to guide the weights during a lift. The weight stack could be adjusted from 18.14 kg to 90.72 kg in increments of 4.54 kg. ILM measures used in the present study were:
>
>> *ILM Curl*: The bar was grasped with palms facing toward the body (underhand grip) and a straight back, bent knee lift was performed to get the bar in position for the curl. The participant then flexed his or her arms to achieve an elbow angle of 90º and raised the bar to elbow height. After 3 warm-up repetitions with approximately 25% of body weight, each participant attempted to curl 50% of his or her body weight. If successful, the weight was increased by 4.54 kg and a second curl was attempted. If unsuccessful, the weight was decreased 4.54 kg, and a second curl was attempted. Each curl attempt was followed by a weight change until the maximum weight that could be curled was determined.
>>
>> *ILM Press* After the ILM curl, participants rested 5 min while the ILM press was explained and demonstrated. For this lift, the lift bar was grasped with the palms of the hands facing away from the body (overhand grip). A straight-

back, bent-knee lift followed by a partial arm extension was used to raise the bar to a 152.4-cm marker on the apparatus. Warm-up lifts and procedures for determining the maximum weight the individual could lift were the same as those for the ILM curl. A lift was disqualified if the individual used his or her legs during the arm portion of the lift, used unsafe lifting techniques, or paused for longer than 1 s during any portion of the lift.

Dynamometer Measurements. A Chatillon Push/Pull mechanical force gauge (Model TCG-250, John Chatillon & Sons, New York, NY) was used to obtain static strength measurements.

Arm Pull. An arm pull test was performed by standing with one hand holding a pull bar attached to the gauge. The other hand was braced against a vertical support to which the gauge was anchored. Feet and toes did not touch the support. The participant then exerted a smooth pull on the handle, generating as much force as possible. A series of 6 pulls was performed alternating the left and right hands. The arm pull score was the average of the pounds of force generated during pulls 3 through 6.

Arm Lift. The arm lift test involved lifting a bar attached to the Chatillon gauge by a chain and cable. The gauge point was set initially at a point equal to the weight of the bar and gauge. The participant stood with feet slightly apart, straddling the cable attaching. Chain length was adjusted so that the bar could be held with the forearms parallel to the floor or angled slightly (i.e., <10°) downward. Participants were instructed to exert the maximum lifting force that they could generate, with back and legs straight, heels flat, and shoulders motionless. The lift was repeated 3 times. The score was based on the last 2 lifts.

Universal Gym Strength Measures. Strength measurements for the arm curl, lat pull-down, shoulder press, and bench press were conducted on a Universal Gym machine. Each measurement started at a resistance determined by the subject's weight. For example, individuals weighing between 160 lb and 189 lb started the bench press with a weight of 110 lb. The weight for each subsequent repetition was increased by 20 lb if the most recent lift appeared to be easy for the subjects or by 10 lb if they appeared to be approaching their limit. Strength was the heaviest weight lifted. This weight was typically reached within 3 or 4 lifts. Procedures for specific tests were:

Arm Curl. Subject stood 6 in to 12 in from the gym with feet shoulder-width apart. The bars were gripped with palms up at approximately shoulder width. Standing erect, with arms fully extended downward, the subject braced his or her arms against the front of the body and lifted the bar to his or her chest. The body could be leaned slightly backward at the start of the lift for bracing, but subjects were not permitted to change this position by leaning backward or forward during the lift.

Bench Press. Subjects laid on their backs on a bench with feet flat on the floor. Position was adjusted so the bar was between the shoulder and nipple line. Handles were gripped at a comfortable width, 1 to 2 handwidths from the shoulder. The bar was pressed to full extension without lifting hips off the bench or feet off the floor.

Leg Press. Subjects sat in the seat provided for this exercise with their lower back against the cushion. The balls of their feet were placed directly over the pedal creases. A goniometer was used to adjust the seat to create a 90º angle, after which the seat was moved forward one notch. The subject gripped the handles on the side of the seat and pushed the pedals to nearly full extension of the legs without locking the knees. The extension had to be performed without lifting up from the seat.

Latissimus (Lat) Pull-Down. Subjects stood facing the machine and grasping the bar handles. The widest possible grip was recommended. Subjects knelt down in front of the machine with torso vertical and elbows fully extended. The bar was pulled to the base of the neck. No movement of the hips or knee joints was allowed. A successful lift was counted if the bar was pulled below the earlobe level.

Shoulder Press. Sitting in front of the machine with the handles of the bar just in front of the shoulder, the subjects assumed a position with back erect and feet on the rung of the stool. The test administrator then adjusted the height of the handles to the level of the shoulders. The weight was pressed to full extension of the arms without leaning back or lifting the feet from the stool rung.

Wingate Tests. Wingate tests were conducted on a cycle ergometer (Monark, Sweden). A metronome was used to control the pedaling rate.

Leg Wingate. Subjects warmed up by pedaling at a rate of 60 rpm for 3 min at a resistance of 1.5 kg. Three 5-s all-out sprints were performed during the second minute of warm-up. During the actual test, the subject began pedaling against no resistance and was instructed to gradually increase the pedaling speed. When the pedaling rate reached 120 rpm, the subject was instructed to pedal as fast as possible. A predetermined resistance based on the subject's weight was applied when the pedaling rate reached 150 rpm. The subject then pedaled as hard and fast as possible for 30 s. The performance measure was the mean power generated during the bout.

Arm Wingate. A Monark arm ergometer was used in this test. Subjects were instructed to kneel behind the ergometer, which had been clamped in place. The subjects then cranked the handles as rapidly as possible for 30 s. The performance measure was the mean power generated during the bout.

Cardiorespiratory Endurance Assessments. The cardiorespiratory element of endurance was measured by three measures. Two measures were derived from a laboratory assessment of maximal oxygen uptake. These measures were the maximal oxygen uptake ($\dot{V}O_{2\max}$) and the anaerobic threshold ($\dot{V}O_{2AT}$). The third test was a timed 1.5-mi run.

$\dot{V}O_{2\max}$. A continuous treadmill protocol was employed to assess $\dot{\mathrm{V}}_{\mathrm{O2\,max}}$. This protocol began with a 2-min walk at 3.0 mph and 0% grade, followed by a 3-min jog at 5.0 mph for women or 5.5 mph for men at 0% grade. From the 6th min through the 17th min, the grade was increased by 2% each minute. If the run lasted longer than 17 min, the grade was held constant at 24%. The speed was increased 0.5 mph at the 18th min and every minute thereafter. Oxygen uptake was measured by open-circuit spirometry.

$\dot{V}O_{2AT}$. This variable was defined as a sharp rise in the ventilatory equivalent of oxygen (i.e., $\dot{V}E/\dot{V}O_2$) relative to oxygen uptake accompanied by a respiratory exchange ratio (RER) close to 1.00. $\dot{V}O_{2AT}$ was the oxygen uptake rate in ml/kg per minute at the inflection point in the plot.

1.5-mi Run. The 1.5-mi run was conducted on a measured, level asphalt track. A 0.25-mi walk/jog warm-up was followed by a brief rest. Subjects then ran 1.5 mi by completing 6 laps on the 0.25-mi track in groups of 2 to 10 persons. Elapsed time was given at each 0.25-mi point in the run. As the subject passed the starting point on each lap, he or she called out his or her name, and the test administrator marked the lap off as completed for that subject. Final times were accepted only when the subject had five marks prior to the completion of the last lap of the test. Completion time recorded to the nearest 0.1 s was the performance measure.

Field Tests for Physical Fitness. Push-ups, pull-ups, sit-ups, broad jump, vertical jump and reach, and a 100-m sprint were performed as additional physical capacity measures.

Push-ups. Push-ups were performed with a partner. The test began by determining the proper location for the partner's fist. That location defined the down position for each push-up. The location was determined by having the subject assume a down position, with hands flat on the floor approximately shoulder-width apart and elbows flexed. The subject then raised him- or herself high enough so the partner could just insert his or her fist upright on the ground touching the participant's shoulder/upper chest region. The test began with subjects in the up position, with hands about shoulder-width apart. The arms, buttocks, and legs were kept straight from head to heels throughout the test. A push-up consisted of lowering the rigid body by flexing the elbows until the subject's shoulder/upper chest touched the partner's fist. The elbows then were extended to return the subject to the up position with arms were fully extended. This sequence was repeated as many times as possible in 1 min, with rest as needed. The partner counted the number of repetitions, while a test administrator counted the number of incorrect repetitions. The score was the total number of push-ups minus the number of incorrect repetitions.

Pull-ups. Subjects faced a pull-up bar, jumped up, and grasped it with the palms of their hands facing toward their bodies. Subjects then hung with arms fully extended and feet off the floor. Using arms and shoulders only, subjects then pulled themselves up until their Adam's apple reached bar level, after which they returned to the lowered position with their arms fully extended. This sequence counted as 1 pull-up. Participants were instructed to perform as many continuous repetitions as they could. The test was terminated as soon as the subject paused for 1 s or longer. Pull-ups were not counted if the subject kicked, swung, or kipped. Pull-ups were not counted if the participant failed to bring his or her Adam's apple to bar height, regardless of where his or her chin was.

Sit-ups. Participants laid on their backs on the floor with knees bent and heels approximately 10 in from buttocks. Arms were crossed over the chest, with the right hand grasping the left shoulder and the left hand grasping the right wrist. A partner held the participant's feet flat on the floor while the participant tightened his or her abdominal muscles to bring the upper body toward the knees. A sit-up was completed when the participant's elbows touched his or her knees and he or she returned to a position with the lower border of the shoulder blades touching the floor. The subject's partner counted the number of sit-ups, and a test monitor kept track of the number of incorrect repetitions. The score was the number of acceptable pushups (i.e., total count – incorrect repetitions) performed in 2 min.

Standing Long Jump. Subjects stood with their toes even with a line that was the zero mark for the jump, feet shoulder-width apart. Subjects then crouched with knees bent and arms swung back. Subjects then jumped by extending the knees and swinging arms forward. Distance was measured from the starting line to the body part touching the jump surface closest to the starting line. Participants were allowed to practice until they achieved good technique, after which they repeated the jump 3 times. The long jump score was the longest of the 3 trial distances.

Vertical Jump and Reach. The subject stood on the test apparatus, with feet approximately shoulder-width apart. The end of a tape measure ribbon was pinned to the right lower leg of the participant's gym shorts. The ribbon was passed through a flattened wire loop and the participant's position was adjusted so that the tape was vertical. The distance from the point on the participant's gym shorts to the flattened wire loop was measured by reading the value where the flattened loop crossed the tape. The participant then crouched, swung his or her arms backward, and bent his or her knees. After a pause, the participant jumped upward as high as possible, swinging arms forward and upward to reach for the highest point possible. The tape was pulled through the flattened loop by the jump. The tape reading at the flattened loop was determined after the jump. The height of the jump was the difference between this post-jump reading and the initial reading. The procedure was repeated 3 times, and the score was the greatest distance of those trials.

100-m Sprint. A 100-m distance was measured off. Subjects stood with their toes behind a starting line that was placed at one end of the course. The test administrator stood at the other end of the course with a stopwatch held above his or her head. The test administrator dropped his or her arm to signal the subject to begin running. The subject ran as fast as possible through the finish line. The test administrator stopped the stopwatch when the subject crossed the finish line. The time showing on the stopwatch was recorded as the test score.

Performance Tasks

Performance measures consisted of two lifting tasks and a carrying task (Beckett & Hodgdon, 1987):

Box Lift to Elbow Height. A box was lifted from the floor to an elbow height platform using a bent-knee, straight-back, two-handed lifting procedure. The box was 33 x 25 x 20 cm, with solid bar handles (20 cm in length x 3.3 cm in diameter), and a weight 5.67 kg when empty. Platform height was determined individually for each subject by determining the height of the bottom of the empty box when the person stood with arms straight and feet shoulder-width apart.

Measurements began with the box loaded to a weight equal to approximately 30% of the participant's body weight for 5 warm-up lifts. The box then was loaded to approximately 60% of body weight. If the participant lifted the box successfully to the platform, the weight was increased 11.34 kg for the next attempt. The procedure was repeated until an unsuccessful lift occurred, after which the weight was decreased to the last successful lift plus 4.54 kg. Lifts then continued, with 4.54-kg increments until another unsuccessful lift occurred. The weight then was set at the last successful lift plus 2.27 kg and a final lift attempt was made. The platform height corresponding to a 90º elbow flexion was determined for each participant.

Box Lift to Knuckle Height. Participants rested for 5 min after the first lifting task. The same box then was lifted from the floor to the previously determined platform height. The starting weight for the lift was the maximum weight lifted to elbow height plus 11.34 kg. Weights then were increased 11.34, 22.68, or 34.02 kg on each subsequent lift, with the amount added based on the test administrator's estimate of how much more the person could comfortably lift. A 1-min rest was taken between lifts, and procedures equivalent to those used in the elbow-height lift were used to determine maximal lifting capacity to within 2.27 kg.

Box Carry. Participants carried a small metal box (33 x 25 x 20 cm) loaded to 34 kg from one platform to another 51.4-m away. Platform heights were set so the box handles were at the height the individual would be holding the box with arms fully extended and feet shoulder-width apart. Participants moved the box from one platform to the other by walking as fast as possible carrying the box in front of them in "…the most comfortable position" (Beckett & Hodgdon, 1987, p. 11). After carrying the box to the platform, the participant returned to the first platform empty-handed to get a second box. Elapsed time was announced at the end of each round trip. Each participant performed the task for two 5-min bouts, with the total distance covered as the performance measure for each bout.

Analysis Procedures

Structural equation models (SEMs) for the study were estimated with LISREL 8.5 (Joreskog & Sorbom, 1996). Anderson and Gerbing's (1988) two-step approach was adopted. Measurement models for ability and performance were developed. Ability-performance relationships then were estimated with the measurement models fixed. Conceptually, this two-step procedure reduces the ambiguity of research findings by ensuring that negative results are not merely manifestations of poor measurement models (Meehl, 1990). Also, this approach reduces the risk that a good measurement model will mask poor fit in the substantive model (McDonald & Ho, 2002).

The ability measurement models for men and women were constructed by a series of parallel analyses for men and women treated as separate samples. Model construction proceeded in several steps. First, each ability measure was assigned to one of four hypothesized ability dimensions. Second, exploratory factor analyses demonstrated that each indicator set defined a single common factor and each indicator met a minimum loading criterion ($\geq .30$) on that factor. Third, a separate SEM was constructed for each hypothesized dimension. This step provided maximum likelihood estimates of the factor loadings and the standard deviations of those loadings. This step also provided an estimate of latent trait variances and established that all of the parameters in the model met the recommended minimum standard (i.e., $|t| \geq 2.00$) for retaining parameters in structural models (Joreskog & Sorbom, 1996). The final step in model construction combined the models for the four individual latent traits into a single model. The factor loadings and latent trait variances were fixed at the values estimated in the initial trait-by-trait structural models. With this constraint, the final step only provided estimates of the covariances among the previously defined latent traits.

The final ability measurement model consisted of two sets of factor loadings and latent trait variances and covariances, one set for men and one for women, for the following hypothetical constructs:

Aerobic Capacity (three indicators): $\dot{V}O_{2\max}$, anaerobic threshold, and 1.5-mi run time.

Dynamic Strength (three indicators): push-ups, pull-ups, and sit-ups.

Anaerobic Power (five indicators): broad jump, vertical jump, 100-m sprint, Arm Wingate Test, and Leg Wingate Test.

Static Strength (nine indicators): arm lift, arm pull, arm curl, lat pull-down, shoulder press, bench press, incremental lift curl, and incremental lift press.

The final model also included a correlated error for shoulder press and bench press. Modification indices from the SEMs were the basis for this addition. These indices represent the minimum expected change in the overall fit of the model if constrained parameters were freely estimated (Joreskog & Sorbom, 1996). These indices must be used cautiously because there is a substantial risk that chance will produce some apparently useful modifications when a large number of constrained parameters must be considered (MacCallum, Roznowski, & Necowitz, 1992). To minimize this risk, a parameter constraint was removed only if its modification index (MI) was large for both men and women. The error covariance for the shoulder press with the bench press was the only parameter that met this criterion (men, MI = 11.81; women, MI = 13.91). This finding represented one specification error in 250 parameters (60 factor loadings plus 190 error covariances) that were fixed at zero in the hypothetical model of four physical abilities that guided the analyses. An estimate of the error covariance for shoulder press and bench press was added to the final model. With this addition, the final measurement model consisted of the hypothesized latent traits, their covariances, and one correlated error.

Performance Measurement Model

The performance measurement model had two dimensions. Measurement models could not be developed for these dimensions separately. Only two indicators were available for each hypothesized performance dimension. A minimum of three indicators is required to define a dimension uniquely (Gorsuch, 1983). Therefore, the performance measurement model was defined a priori as 2 correlated dimensions with two tasks loading on each dimension.

Two constraints were imposed to identify the model. First, the scaling of latent traits was established by fixing the variance of those traits at 1.00. Second, the factor loadings for the two indicators defining each trait were constrained to be equal.

Substantive Models

Substantive models quantified the ability-performance associations. Physical abilities were treated as exogenous causal variables in these models. Performance dimensions were treated as endogenous dependent variables.[1] The combined ability and performance dimensions were employed to construct and test a systematic series of explanatory models for performance:

1. Null [N] model: All ability-performance relationships were fixed at 0.00.
2. SS model: SS affected both performance dimensions.
3. SS/DS model: DS effects were added to the SS model.
4. SS/DS/AP model: AP effects were added to the SS/DS model.

[1] The possibility that ability tests and task performance were merely different manifestations of a single set of underlying abilities had been examined in prior studies. Treating ability and performance as the products of a single set of latent traits produced models with poor fit relative to models that treated them as distinct exogenous and endogenous constructs. That alternative therefore was not pursued further here.

5. SS/DS/AP/AC model: AC effects were added to the SS/DS/AP model.

This sequence of models was chosen to address several specific research issues. First, the SS-performance parameters in the SS model could be compared to the SS-performance estimates from equivalent models in prior work (Vickers, 1996, 1997, 2003a). Second, the SS/DS model for the carrying tasks replicated Vickers' (2003a) model for moderate-duration tasks. The two remaining models extended the search for omitted variable bias effects. By adding AC last, the comparison between the final model in the sequence and the preceding model provided an immediate basis for determining whether AC affected performance independent of the other physical abilities.

The final model minimized the risk of omitted variable bias as far as possible in this study. Any substantial ability-performance association in this model would be free of bias from three other established physical ability dimensions. If there were any remaining bias, it would have to come from abilities that were not represented in this study (e.g., quality of movement; Hogan, 1991b).

The SS/DS/AP/AC model was trimmed to establish a final model. The trimming procedure emphasized generalizability across men and women. Pooled significance tests (Rosenthal, 1978) were applied to identify replicated misfit. A Bonferroni significance criterion ($p < .05/8 = .00625$; cf. Green, Thompson, & Poirer, 2001) was used to control for the risk of capitalizing on chance (MacCallum et al., 1992). This approach emphasized model parsimony. The relatively extreme significance criterion reduced the power of significance tests thereby increasing the likelihood that small effects will be eliminated from the model.

Model comparisons followed recommendations that multiple indicators of model adequacy should be used in model selection (Boomsma, 2000; Hu & Bentler, 1998, 1999; McDonald & Ho, 2002). In the present case, the standardized root mean square residual (SRMR; Joreskog & Sorbom, 1996 and the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993; Steiger, 1990) were chosen as indices that are sensitive to model misspecification (Fan, Thompson & Wang, 1999; Hu & Bentler, 1998). The nonnormed fit index (NNFI, Bentler & Bonett, 1980; Tucker & Lewis, 1973) was chosen to represent SEM analogues of R^2 in regression analysis.

Results

Ability Trait Correlations

The risk of omitted variable bias would disappear if physical ability traits were uncorrelated. Unfortunately, the analysis produced evidence of weak to moderate relationships between traits (Table 1):

- Ability trait correlations were approximately equal for men and women. None of the differences were statistically significant using a two-tailed test ($|z| < 1.79$, $p > .074$).

Table 1. Latent Trait Correlations for Ability

	Aerobic Capacity (AC)		Dynamic Strength (DS)		Anaerobic Power (AP)		Static Strength (SS)	
	M	W	M	W	M	W	M	W
AC	1.000	1.000						
DS	.448	.706	1.000	1.000				
AP	.305	.444	.408	.664	1.000	1.000		
SS	.164	.204	.566	.593	.565	.545	1.000	1.000

Note. *N* = 36 for women; N = 55 for men.

Thus, the average of the correlations for men and women provided a reasonable summary of the data for both genders.

- Four of six ability correlations were moderately large (SS-DS, average *r* = .577; SS-AP, AC-DS, average *r* = .562; average *r* = .556; DS-AP, average *r* = .519). AC was weakly related to SS (average *r* = .180) and moderately related to AP (average *r* = .361).

Clearly, the data displayed evidence of one condition that would raise concerns about omitted variable bias, the presence of substantial correlations between potential causes.

Ability and Performance

A diffuse pattern of associations was evident when the correlations between ability latent traits and performance latent traits were examined. Each of the eight ability-performance correlations met Cohen's (1988) minimum effect size criterion (*r* = .10) when the male and female results were averaged (cf., Table 2). All gender differences were statistically nonsignificant ($|z| \leq 0.67$, *p* > .503). Despite the broad tendency toward positive correlations, the patterns of association were noticeably different for carrying and lifting.

Table 2. Ability-Performance Latent Trait Correlations

	Carrying		Lifting	
	M	W	M	W
Aerobic Capacity (AC)	.508	.423	.104	.110
Dynamic Strength (DS)	.385	.443	.361	.320
Anaerobic Power (AP)	.320	.334	.360	.294
Static Strength (SS)	.394	.394	.637	.540

Note. *N* = 36 for women; *N* = 55 for men.

Table 3. Summary of Ability-Performance Models

Model	χ^2	*df*	$\Delta\chi^2$	NCP	F_0	RMSEA	SRMR	NNFI
Men								
Null	160.91	80		80.91	1.471	.136	.432	
SS	108.59	78	52.32	30.59	.556	.084	.185	.612
SS/DS	100.13	76	8.46	24.13	.439	.076	.165	.686
SS/DS/AP	99.50	74	0.63	25.50	.464	.079	.162	.659
SS/DS/AP/AC	79.96	72	19.54	7.96	.145	.045	.143	.891
Trimmed[a]	81.73	77	(1.77)[b]	4.73	.086	.033	.147	.939
Women								
Null	112.88	80		32.88	.913	.107	.379	
SS	97.38	78	15.50	19.38	.538	.083	.206	.395
SS/DS	90.81	76	6.57	14.81	.413	.074	.166	.526
SS/DS/AP	87.72	74	3.09	13.72	.381	.072	.159	.549
SS/DS/AP/AC	76.74	72	10.98	4.74	.132	.043	.148	.840
Trimmed[a]	81.34	77	(4.60)[b]	4.34	.121	.040	.158	.863

Note. Null χ^2 = Observed Null Model – Sum of Measurement Models.
[a] Deleted DS and AP effects plus effect of AC on Lifting.
[b] Parentheses indicate χ^2 increase from the preceding model.

- *Lifting*. SS, the strongest predictor (average r = .601), accounted for 3 to 4 times as much variance as DS (average r = .345) or AP (average r = .335). AC effects were weak (average r = .106; pooled z = 0.98, p >.163, two-tailed).
- *Carrying*. All of the ability dimensions were moderately related to lifting performance (AC, average r = .476; DS, average r = .408; AP, average r = .326; SS, average r = .394).

The second condition for omitted variable bias was satisfied. Each performance dimension was related to several of the correlated ability dimensions.

Structural models. Table 3 presents the planned series of ability-performance model comparisons plus the final model obtained by trimming nonsignificant parameters from the SS/DS/AP/AC model. The following statements describe the general effects of ability with the two performance dimensions considered together for each model:

- Model 1: SS predicted performance for men ($\Delta\chi^2$ = 52.35, 2 *df*, p < .001) and women ($\Delta\chi^2$ = 15.50, 2 *df*, p < .001). The combined effect, which is the sum of the separate effects, was significant ($\Delta\chi^2$ = 67.85, 4 *df*, p < .001).
- Model 2: Adding DS significantly improved the model for men ($\Delta\chi^2$ = 8.46, 2 *df*, p < .015) and women ($\Delta\chi^2$ = 6.57, 2 *df*, p < 038). The combined effect was significant ($\Delta\chi^2$ = 15.03, 4 *df*, p < .005).

Table 4. Path Coefficients for Ability Effects on Performance

	Carrying		Lifting	
	Men	Women	Men	Women
Aerobic Capacity (AC)	.456	.357		
Static Strength (SS)	.319	.322	.637	.540
R^2	.357	.278	.406	.292

Note. Table reports standardized path coefficients. SS-AC correlations were modest (men, *r* = .164, *N* = 55; women, *r* = .204, *N* = 36).

- Model 3: Adding AP did not improve the fit for men for men ($\Delta\chi^2$ = 0.72, 2 *df*, *p* > .697) or women ($\Delta\chi^2$ = 3.76, 2 *df*, *p* > .152). The combined effect was not significant ($\Delta\chi^2$ = 4.48, 4 *df*, *p* > .344).
- Model 4: Adding AC to the SS/DS/AP model significantly improved the fit of the model for men ($\Delta\chi^2$ = 19.44, 2 *df*, *p* > .001) and women ($\Delta\chi^2$ = 10.31, 2 *df*, *p* > .001). The combined effect was significant ($\Delta\chi^2$ = 29.75, 4 *df*, *p* < .001).
- Model 5: Dropping both effects for DS and AP and the AC-Lifting effect had little effect on model fit. The increase in misfit was not significant for men ($\Delta\chi^2$ = 1.76, 3 *df*, *p* > .623), for women ($\Delta\chi^2$ = 4.59, 3 *df*, *p* > .204), or for both sexes together ($\Delta\chi^2$ = 6.35, 6 *df*, *p* > .385)
- Trimmed model: The trimmed model was the SS model with an effect added to reflect the impact of AC on Carrying. Adding the AC-Carrying effect to the SS model improved the fit of the model for men ($\Delta\chi^2$ = 26.86, 1 *df*, *p* < .001), for women ($\Delta\chi^2$ = 16.05, 1 *df*, *p* < .001), and for both sexes together ($\Delta\chi^2$ = 42.91, 2 *df*, *p* < .001)
- SS/AC trimmed model: This model did not provide the best absolute fit to the data. Other models that were considered provided better fit as indicated by the parentheses in Table 3. Trimming slightly increased the misfit between the model and the data, but the increase was small relative to the number of parameters eliminated. This result was expected because the fit of models will almost always improve when additional parameters are introduced (Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989). At the same time, it is often the case that some elements of a model contribute little to its overall accuracy. These statistical facts made the inspection of other criteria particularly important for this model selection process. That inspection showed that:
    - RMSEA was smallest for the trimmed model for both men and women. In both cases, RMSE was <05, the recommended criterion for accepting a model as having adequate fit (Browne & Cudeck, 1993).
    - NNFI was largest for the trimmed model. For men, the NNFI of .940 exceeded the .900 criterion recommended by Bentler and Bonett (1980) . For women, the NNFI of .863 approached this criterion value.
    - SRMR was smallest for the trimmed model for both men and women.

Thus, the trimmed model was the best of the five alternatives by all four criteria for both men and women.

Path coefficients in the trimmed model were comparable for men and women (Table 4). AC and SS had comparable effects on Carrying performance. The SS/AC model's explanatory power was modest (i.e., 27% to 41% of the performance variance).

Replication of SS/DS Bias Effects

The SS/DS model replicated Vickers' (2003a) findings for moderate duration tasks. The earlier study indicated that both SS and DS affected performance in a predominantly male sample. This result replicated in the present data; adding an effect of DS on Carrying to the SS model improved the fit for men ($\chi^2 = 4.95$). The DS effect on Carrying was consistent with prior findings (*b* = .018, *t* = 1.79). This effect was not significant in this sample, but the combined results would have been significant if this estimate had been pooled with Vickers' (2003a) earlier findings. However, the present analyses showed that this model did not generalize to women in either magnitude ($\chi^2 = 0.47$) or sign (*b* = -.006, *t* = -.57).

Gender-Specific Models

The modification indices for the potential associations of latent traits that had been excluded from the model were examined. The goal was to identify any latent trait associations that were specific to either men or women. Every index was small (men, MI < .50; women, MI < 1.76). The associations in the trimmed model therefore were judged both necessary and sufficient for both men and women.

Search for Ability-Task Specificity

Specific physical abilities might be critical to specific tasks. If so, an ability-performance model limited to general dimensions would tell only part of the story. Standardized residuals were examined to evaluate this possibility. The residuals from male and female models were treated as replications in this examination. This treatment was chosen to minimize the risk of capitalizing on chance (MacCallum et al., 1992). This treatment also was justified to some extent by the general similarity of the findings for men and women up to this point in the analysis. The findings were:

- Residuals were normally distributed (women, Kolmogorov-Smirnov *Z* = .50, *p* = .966; men, K-S *Z* = 1.20, *p* = .112).
- No individual residual met Green et al.'s (2001) stepwise Bonferroni criterion (p < .05/160 = .0003125, *z* = 3.42). In fact, only the Arm Wingate Test – Box Lift to Elbow Height residual even approached this value (z = 3.37). All other residuals were considerably smaller ($|z| < 3.00$ for all).
- Male and female residuals were weakly correlated (*r* = .346). This trend suggested that residuals showed a weak general tendency to replicate across genders. Pooled probabilities were computed by the method of adding *p*s (Rosenthal, 1978). This computation was undertaken to determine whether this general tendency for male residuals to correspond to female residuals included any specific residuals that were significant when pooled across genders. No residual met the Bonferroni criterion (*p* < .00063). In fact, only four were significant at *p* < .05 (Box Carry Bout 1 – Arm Wingate

Test, *p* = .0029; Box Carry Bout 2 – Arm Wingate, *p* = .0021; Box Lift to Elbow Height – Arm Wingate Test, *p* = .0441; Box Lift to Knuckle Height – Arm Wingate Test, *p* = .0015).

Taken together, these findings provide some further evidence that the treatment of the Arm Wingate Test affected the fit of the model. However, even if one assumes that the pattern reflects the decision to treat the Arm Wingate Test as an index of AP, the fact that pooled results were not statistically significant indicates that the misfit produced by this decision was too small to justify adding specific task-test associations to the final model.

Discussion

The study results can be examined from two related perspectives. One perspective focuses on omitted variable bias to help define a general issue for physical ability-task performance modeling. The other perspective focuses on providing a substantive model of the physical ability determinants of lifting and carrying performance. These perspectives are inextricably interrelated because the general issue must be dealt with to ensure that the substantive model has meaning.

Researchers should be sensitive to the risk of omitted variable bias when modeling the association of physical ability with performance. Both preconditions for omitted variable bias were clearly met in this study. Ability dimensions were correlated. The present correlations were moderate in magnitude, but stronger relationships have been observed in other studies (Myers et al., 1993). In addition, there was a broad general tendency for all abilities to be related to both elements of performance. In particular, seven of eight bivariate ability-performance correlations were significant. Despite this diffuse pattern of associations, the final model included only three effects of ability on performance, SS-Lifting, SS-Carrying, and AC-Carrying. The other four significant bivariate associations illustrate the potential for developing biased models. Studied in isolation, each bivariate relationship could be the basis for a causal model with significant predictive accuracy. However, any model that included a causal effect of DS or AP on either lifting or carrying would be based on omitted variable bias. The same may be true of including an effect of AC on lifting performance. The risk of bias is real and substantial.

The findings also illustrated that omitted variable bias can be difficult to eliminate from ability-performance models. Simple replication of empirical associations is not sufficient to rule out bias. Even incomplete multivariate analyses will not rule out this bias. Vickers' (2003a) SS/DS model for men replicated in the present analyses. The DS effect was not included in the final model, so this association was an instance of replicated omitted variable bias. This replication should not be surprising. If two studies produce the same correlations among abilities and between abilities and performance, the models derived from those correlations will replicate. The replication depends on the pattern of bivariate associations, not the truth of the model. The obvious implication is that model replication is not equivalent to model validation. Validity must be established by ruling out plausible alternative models. A thorough search for potential bias is required to achieve this goal.

The fact that replication does not rule out omitted variable bias directs attention to the requirements for dealing with this problem. To avoid omitted variable bias, researchers must either include all relevant causal variables in their models or ensure that omitted variables are not correlated with variables that have been included (James et al., 1982). When considering physical abilities as causes of task performance, the fact that physical abilities are moderately (cf., Table 2) to strongly (Myers et al., 1993) correlated is very important. The first option of simplifying the research problem by omitting some ability variables is not viable. The only available option is to include all ability dimensions in the model as potential causal factors for physical performance.

The situation is not as bleak as the preceding conclusion might make it appear. In the context of physical ability-task performance models, the causal variables appear to be limited to general physical ability dimensions. This inference is based on the lack of replicable residuals in the studies to date. The conclusion from this line of reasoning is that a well-defined general model of physical abilities is a critical requirement for eliminating omitted variable bias in physical ability-performance models. Factor analytic studies of physical abilities suggest that 3 to 7 general dimensions must be measured to adequately represent this domain (e.g., Fleishman, 1964; Hogan, 1991b; Myers et al., 1993).[2] Further research to better define the physical ability domain and identify the best marker variables for each latent ability trait in that domain would be helpful for future studies of physical performance.

The final model in this study indicated that strength and aerobic capacity were sufficient to represent ability in these data. Logic suggests that this simple model is likely to generalize to other settings, but further exploration along two lines would be worthwhile. First, studies should be undertaken to evaluate the effects of ability dimensions that have been omitted from existing studies. For example, Hogan's (1991b) physical ability model included balance and flexibility dimensions, both of which were absent from the present study. These dimensions may not affect performance, but the possibility should be investigated. Second, the structure of performance should be explored further. The tasks in this study simulated typical U.S. Navy manual material-handling activities (Beckett & Hodgdon, 1987). The ability requirements for performance may be different if a wider range of tasks were investigated. It should be noted, however, that prior research indicates the present findings are likely to generalize to a wide range of manual material-handling tasks (Vickers, 1995, 1996, 2003a).

This study helped clarify the boundaries of physical ability-task performance models, but it also increased the uncertainty about some details of the model. The association between general strength and performance on brief manual material-handling tasks has now been investigated four times. Standardized effect size estimates in those studies were $r = .742$ for lifting and $r = .962$ for brief (i.e., <1 min) carrying (Vickers, 1995) tasks, $r = .962$ for an overall task performance (Vickers, 1996), and $r = .86$ for tasks lasting several minutes (Vickers, 2003a). The present values for lifting ($r = .637$ for men; $r = .540$ for women) and moderate duration carrying (men, $r = .319$; women, $r = .322$) extended this range of effects downward.

[2] The upper limit of this range may be too high. That limit is based on Fleishman's (1964) findings. Recent simulation studies (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Lautenschlager, 1989) give reason to believe the factor extraction criterion was too lenient. Four factors might have sufficed (cf., Myers et al., 1993).

Methodological factors, such as the screening procedures and the limited range of tasks in this study, may have contributed to the weaker associations in this sample.

The results have selection, training, and job design implications. With regard to selection, the findings are consistent with the strength/aerobic capacity schema developed by Vogel, Wright, Patton, Dawson, and Eschenback (1980) to classify U.S. Army occupations. The findings also support the U.S. Air Force's use of strength criteria as occupational requirements (Ayoub, Jiang, Smith, Selan, & McDaniel, 1987). The simplicity of the schema is important. Any physical selection criteria that are added to existing selection profiles can be expected to reduce the pool of qualified applicants for an occupation (Marston, Kubala, & Kraemer, 1981). This undeniable effect is less problematic if only a few ability dimensions must be considered.

The ability-performance model has straightforward training implications. Physical training programs should be designed to develop both strength and aerobic capacity. The fact that training can greatly increase both abilities (Rhea, Alvar, Burkett, & Ball, 2003; Londeree, 1997; Wolfe, LeMura, & Cole, 2004) could eliminate the potential selection bottleneck imposed by adding physical ability to screening profiles.

Omitted variable bias is important when considering the training implications of the final ability-performance model. Physical training programs often emphasize the DS dimension. Significant bivariate associations between dynamic strength markers (e.g., push-ups) and performance are one justification for this practice. The current findings indicate that these bivariate associations are less likely to represent causal effects than they are to represent omitted variable bias. According to the present model, training that enhances DS will not improve performance. This point exemplifies the practical implications of omitted variable bias.

This study has noteworthy limitations. The samples were small and selected by screening on strength. These study characteristics should tend to offset one another. Correlations tend to be overestimated in small samples (Edwards, 1984). Restriction of range leads to underestimation (Sackett & Yang, 2000). The net effect of these trends in the present case is uncertain.

Sample size also affects significance tests. Small samples may have simplified the model by reducing small, but potentially important, effects to statistical insignificance. However, model selection included other criteria in addition to significance tests. This inclusion should help control the effects of small sample size for the latent trait components of the model. However, significance tests were the sole basis for evaluating residuals. In this case, consideration of additional contextual factors gives reason to believe that small sample size was not the key to model simplification. To begin with, residuals were normally distributed for men and for women. This pattern of residuals would be expected if these statistics were produced by chance. Also, large residuals did not replicate across genders in either this study or an earlier one (Vickers, 1996). These points must be considered in light of the fact that models are not really expected to account for the full complexity of behavioral phenomena (MacCallum, 2003). The question is not whether the model is literally true in the sense of accounting for all systematic trends in the data. Instead, the central question is whether the model is close enough to be acceptable (Serlin & Lapsley, 1985). From this perspective, the available evidence supports the view that models that rely on higher-order factors as explanatory variables adequately account for the test-task covariation pattern. Systematic searching for points at which this approximation fails could be constructive. For example, grip strength can be isolated as a narrow physical ability facet (Vickers, 2003b) and could logically be essential for tasks such as stretcher carrying. However, such explorations should be undertaken with the understanding that Knapik, Harper, Crowell, Leiter, and Mull's (1998) findings may be a typical outcome of such searches. In that study, grip strength was related to stretcher-carrying performance, but lat pull-down strength was a slightly better predictor and bench press strength was nearly as good a predictor. An association between stretcher carrying and SS or some related general strength indicator would be one explanation for these results.

In conclusion, meaningful models of the effects of physical ability on task performance must consider the risk of omitted variable bias. Proper treatment of this problem is likely to produce simpler models. The available evidence supports the view that general strength and aerobic capacity are the critical abilities for physical task performance. This view is supported by the general body of research within the domain of exercise physiology (McArdle, Katch, & Katch, 2001). This conclusion is tempered by the fact that the available models are based on data from studies that were not guided by definitive measurement models for either physical abilities or task performance. Also, available studies did not involve a systematic search for points at which this simple model might break down. The benefits of pursuing this model include improved personnel selection and appropriately focused physical training.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423.

Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology, 67,* 588-604.

Astrand, P. O., & Rodahl, K. (1986). *Textbook of work physiology* (3rd ed.). New York: McGraw-Hill.

Ayoub, M. M., Jiang, B. C., Smith, J. L., Selan, J. L., & McDaniel, J. W. (1987). Establishing a physical criterion for assigning personnel to U.S. Air Force jobs. *Am Ind Hyg Assoc J, 48*(5), 464-470.

Beckett, M. B., & Hodgdon, J. A. (1987). *Lifting and carrying capacities relative to physical fitness measures* (NHRC Technical Report No. 87-26). San Diego, CA: Naval Health Research Center.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Boomsma, A. (2000). Reporting analysis of covariance structures. *Structural Equation Modeling, 7*(3), 461-483.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53,* 585-596.

Edwards, A. L. (1984). *An introduction to linear regression and correlation* (2nd ed.). New York: W. H. Freeman and Company.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), 56-83.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice-Hall.

Green, S. B., Thompson, M. S., & Poirer, J. (2001). An adjusted Bonferroni method for elimination of parameters in specification addition searches. *Structural Equation Modeling, 8*(1), 18-39.

Hogan, J. C. (1991a). Physical Abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 753-831). Palo Alto: Psychologists Press.

Hogan, J. (1991b). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76*(495-507).

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage Publications.

Joreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Knapik, J. J., Harper, W. H., Crowell, H. P., Leiter, K. L., & Mull, B. T. (1998). *Standard and alternate methods of stretcher carriage: performance, human factors, and cardiorespiratory response* (Technical Report ARL-TR-1596). Aberdeen Proving Ground, MD: Army Research Laboratory.

Kroemer KHE, Kroemer HJ, Kroemer-Elbert KE. 1990. *Engineering physiology: bases of human factors/ergonomics*. New York: Van Nostrand Reinhold

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research, 24*(3), 365-395.

Londeree, B. (1997). Effect of training on lactate/ventilatory thresholds: A meta analysis. *Med Sci Sports Exerc, 29*, 837-843.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

Marston, P. T., Kubala, A. L., & Kraemer, A. J. (1981). *The impact of adopting physical fitness standards on Army personnel assignment: A preliminary study* (HUMRRO-FR-MTRD-TX-80-6). Alexandria, VA: Human Resources Research Organization.

McArdle, W. D., Katch, F. I., & Katch, V. L. (2001). *Exercise physiology: Energy, nutrition, and human performance* (5th ed.). Philadelphia: Lippincott Williams and Wilkins.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141.

Monod, H. (1985). Contractility of muscle during prolonged static and repetitive dynamic activity. *Ergonomics, 28*(1), 81-89.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430-445.

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The dimensions of human physical performance: Factor analyses of strength, stamina, flexibility, and body composition measures. *Human Performance, 6*(4), 309-344.

Rhea, M. R., Alvar, B. A., & Burkett, L. N., & Ball, S. D. (2003). A meta-analysis to determine the dose response for strength development. *Medicine and Science in Sports and Exercise*, *35*, 456-464.

Robertson, D. W., & Trent, T. T. (1985). *Documentation of muscularly demanding job tasks and validation of an occupational strength test battery (STB)* (NPRDC Technical Report No. 86-1). San Diego, CA: Navy Personnel Research and Development Center.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*(1), 185-193.

Sackett, P. R., & Yang, H. (2000). Correction for restriction of range: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73-83.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika,* 38,1-10

Vickers, R. R., Jr. (1995). *Physical task performance: Complexity of the ability-performance interface* (NHRC Technical Report No. 95-30). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (1996). *Generalizability test of a physical ability-job performance model* (NHRC Technical Report No. 96-16). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2003a). *Physical strength and the performance of moderate duration tasks* (NHRC Technical Report No. 03-08). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2003b). *The measurement structure of strength* (NHRC Technical Report No. 03-30). San Diego, CA: Naval Health Research Center.

Vogel, J. A., Wright, J. E., Patton, J., F., Dawson, J., & Eschenback, M. P. (1980). *A system for establishing occupationally-related gender-free physical fitness standards* (USARIEM Technical Report). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

Wolfe, B. L., LeMura, L. M., & Cole, P. J. (2004). Quantitative analysis of single- vs. multiple-set programs in resistance training. *Journal of Strength and Conditioning Research, 18,* 35-47.

Appendix A. Development of Measurement Models

The following procedures were employed to develop the measurement models for ability and performance. The steps for the ability model were:

- Ability tests were classified a priori as indicators for one of four hypothesized ability dimensions, Static Strength (SS), Dynamic Strength (DS), Anaerobic Power (AP), or Aerobic Capacity (AC) as described on p. 9.
- Principal factors analyses (PAF; SPSS, Inc., Chicago, IL) tested the claim that each set of indicators was unidimensional. The analyses produced a single common factor for each set of variables. Every factor loading was large enough to treat the variable as an acceptable indicator of the hypothesized construct (i.e., >.30, absolute).
- Confirmatory factor analysis (CFA) produced latent trait loadings for each set of indicators. Separate analyses were carried out for each of the four hypothesized ability dimensions. Each model was treated as a single latent trait. The scale of the latent trait was established by fixing the factor loading at 1.00 for one of the indicator variables. Every *t* value in these four analyses exceeded Joreskog and Sorbom's (1996) recommended $t \geq 2.00$ criterion. Each subset was analyzed separately to obtain factor loadings that were based solely on the relationships among indicators of the same theoretical construct.
- Results for men and women were treated as a replication. Loadings with *t* values that were slightly less than 2.00 for one group were accepted if the value was well above 2.00 in the other group.
- The four unidimensional CFA models were combined into an overall physical ability model. The latent trait loading for each indicator on the relevant dimension was fixed at the value determined in the prior step. Correlations between the ability traits were estimated to complete the specification of the physical ability measurement model.
- Examination of the complete measurement model in follow-up analyses identified reasons why 1 of the 20 indicators should be moved from its hypothesized dimension to a different dimension:

    - The Arm Wingate Test produced a large MI on the SS dimension for men (MI = 19.86) and for women (MI = 3.84).
    - Extending the measurement model to include a loading for the Arm Wingate Test on the SS dimension improved the fit of the model to the data. At the same time, this modification resulted in large MI values for the Arm Wingate Test on the AP dimension (men, $\chi^2 = 13.29$, women, $\chi^2 = 7.19$). The estimated parameter changes produced by LISREL 8.5 indicated that free estimation of this parameter would substantially reduce the parameter value for men (-.51) and for women (-.29).
    - The constraint on the Arm Wingate Test loading on the AP dimension was removed. When freely estimated in the full ability measurement model, the Arm Wingate Test loading on AP was too small ($|t| < 0.53$) to retain this indicator as part of the AP model.
    - Based on the preceding analyses, the Arm Wingate Test could have been assigned to the SS dimension in the final measurement model. This decision would raise questions about the conceptual interpretation of the AP and SS ability dimensions.

Table A-1. Ability Measurement Model Parameters

	Men			Women		
	λ_x	*t* value	θ_Δ	λ_x	*t* value	θ_Δ
Aerobic Capacity						
$\dot{V}O_{2\max}$	1.000	-	7.764	1.000	-	
$\dot{V}O_{2AT}$	.689	7.28	12.915	.638	9.28	9.365
1.5-mi Run	-.345	-10.00	.509	-.278	-9.59	1.663
Muscle Endurance						
Pull-up	.245	7.86	11.931	.172	3.28	2.552
Push-up	1.000	-	-	1.000	-[a]	29.800
Sit-up	.646	4.31	275.924	1.752	1.752	197.356
Anaerobic Power						
Broad Jump	.0029	2.64	.0123	.0039	3.05	.0098
Vertical Jump	.0008	2.57	.0034	.0011	2.94	.0008
100-m Sprint	-.0143	-2.59	.8984	-.042	-2.87	2.306
Arm Wingate	.420	2.43	6485.0830	.545	2.38	925.227
Leg Wingate	1.000	-	30161.3332	1.000	-	4436.332
General Strength						
Arm Lift	.789	5.25	159.657	.309	2.61	14.859
Arm Pull	.670	6.51	62.865	.362	2.30	26.788
Arm Curl	.540	7.70	23.973	.403	4.17	8.562
Lat Pull-down	1.000	-	67.059	1.000	-	6.653
Shoulder Press	.855	9.05	27.822	.671	4.96	14.744
Bench Press	1.303	8.71	77.367	.729	4.72	19.958
Leg Press	2.459	6.94	707.855	1.797	2.17	742.900
ILM 1	.856	6.48	105.767	.601	3.88	22.787
ILM 2	.880	8.36	45.754	.571	4.35	15.398

Note. Indicators with λ_x = 1.00 and no *t* values were the measures chosen to set the scale for the latent trait. λ_x is the loading for the variable on the latent trait. θ_Δ is the residual for the indicator.

Given that the analyses confirmed 20 of 21 the initial hypothetical assignments of measures to latent traits, the original conceptual model appeared to provide a reasonably robust representation of physical abilities. Additional analyses were carried out with the Arm Wingate assigned to the SS factor. This reassignment had minor effects on bivariate associations within the ability domain and between the ability and performance domains. The reassignment did not affect the form of the final model. Given these results, maintaining the overall correspondence between theoretical constructs and the measurement model was a more reasonable course of action than reassigning the Arm Wingate Test. The decision to leave this test aligned with the latent trait to which it was assigned originally maintained the conceptual integrity of the latent traits at the possible cost of sacrificing some valid variance in one indicator variable. Table A-1 presents the resulting measurement model.

Table A-2. Performance Measurement Model Parameters

	Men			Women		
	λ_y	*t* value	θ_ε	λ_y	*t* value	θ_ε
Carrying						
Bout 1	73.248	9.70	656.921	66.589	7.96	-[a]
Bout 2	73.248	9.70	842.623	66.589	7.96	1144.014
Lifting						
Elbow Height	1.000	-	19.365	1.000	-	14.478
Knuckle Height	1.176	3.51	119.063	1.733	3.78	64.025

[a]The initial variance estimate was negative, so this parameter was fixed at 0.00.

- Each of the final measurement models contained an indicator with a negative residual variance. These residuals were fixed at 0.00 in the measurement models used in the analyses. The decision was based on the assumption that negative variation was impossible, so those parameters represented chance events. Fixing the values at 0.00 was the minimum feasible deviation from the estimated value.

The physical ability measurement model shown in Table A-1 accounted for virtually all of the reliable systematic covariance between different ability tests. After conducting separate analyses for men and women, only six residuals were statistically significant ($p < .05$) for both men and women. Only three of the six replicable residuals produced combined χ^2 values that exceeded the $p < .0003$ Bonferroni criterion for significance given 190 significance tests and $p < .05$ as the experiment-wide error. All three paired the Arm Wingate with a strength measure (ILM2: men, χ^2= 16.64; women, χ^2 = 7.85; Lat Pull-down: men, χ^2 = 14.32; women, χ^2 = 6.95; Arm Curl, men, χ^2 = 15.76; women, χ^2 = 4.79). The combined χ^2 values for these three residuals all markedly exceeded the Bonferroni criterion. None of the other replicable residuals were close to the Bonferroni criterion ($p > .0029$ for all). Thus, the ability measurement model accurately summarized the covariation of the physical ability tests with the exception of some specific associations of the Arm Wingate with selected strength tests. The fact that the three significant replicable associations all involved upper body strength measures might suggest that these discrepancies indicate that the model should include separate latent variables for upper and lower body strength. However, the misfit did not extend to other upper body strength measures in the data, including the Arm Lift, Arm Pull, Shoulder Press, and Bench Press. Furthermore, an extensive analysis of tensiometer strength measures has shown that a distinction between upper and lower body strength could not be justified empirically (Vickers, 2003b).[3]

[3] Specific factors could have been introduced into the model to reflect the significant associations. However, the anticipated gain in goodness of fit would have been modest, particularly for men. Instead, the decision was made to determine whether the paired variables had similar patterns of residuals when related to performance measures. Similar patterns would point to the usefulness of introducing a specific factor common to the pair of tests.

Table A-2 presents the performance measurement model. Note that the weights for carrying bouts were constrained to be equal. This constraint provided the scaling for the carrying dimension.

Table A-3 provides the covariance matrices for the ability measurement model. The latent trait correlations reported in the body of the paper were derived from these matrices.

Table A-3. Covariance Matrices for Latent Traits

	Aerobic Capacity	Muscle Endurance	Anaerobic Power	General Strength
Women				
Aerobic Capacity	56.59			
	(13.53)			
Muscle Endurance	44.07	68.93		
	(6.12)	(28.16)		
Anaerobic Power	126.53	209.02	1455.59	
	(36.95)	(36.87)	(959.64)	
General Strength	9.23	29.59	124.08	36.10
	(7.11)	(6.62)	(28.21)	(10.75)
Men				
Aerobic Capacity	38.20			
	(9.09)			
Muscle Endurance	40.66	227.87		
	(2.30)	(43.85)		
Anaerobic Power	149.36	461.62	5125.17	
	(61.20)	(131.30)	(3813.03)	
General Strength	13.87	114.03	549.54	173.78
	(10.76)	(16.51)	(99.34)	(45.11)

Note. Parameter values appear in the first line. Parenthetical values on the second line are *t* values.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD MM YY)
15 09 08

2. REPORT TYPE
Technical

3. DATES COVERED (from – to)
Jul 08 – Sep 08

4. TITLE AND SUBTITLE
Physical Ability-Task Performance Models: Assessing the Risk of Omitted Variable Bias

5a. Contract Number:
5b. Grant Number:
5c. Program Element Number:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60704

6. AUTHORS
Vickers, Ross R., Jr.; Hodgdon, James A; Beckett, Marcie B.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER
09-04

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Commanding Officer
Naval Medical Research Center
503 Robert Grant Ave
Silver Spring, MD 20910-7500

Commander
Navy Medicine Support Command
P.O. Box 140
Jacksonville, FL 32212-0140

10. SPONSOR/MONITOR'S ACRONYM(S)
NMRC/NMSC

11. SPONSOR/MONITOR'S REPORT NUMBER(s)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The physical capacities of job incumbents limit performance on occupational physical tasks. While muscle strength is logically an important performance-relevant physical ability, omitted variable bias may cause its importance to be overstated. This bias occurs when a causal variable in a model correlates with other causal variables that are omitted from the model. The impact of omitted variable bias on the strength-performance association was evaluated in a study of simulated job performance in men and women. The study measured 4 major abilities, Static Strength (SS), Dynamic Strength (DS), Anaerobic Power (AP), and Aerobic Capacity (AC). Performance measures were simulated lifting and carrying tasks. Analysis showed moderate to strong relationships among the ability measures. All four ability measures were significantly related to lifting and to carrying performance. However, construction of a series of alternative predictive models led to adoption of a final model, with SS and AC as the only predictors. The absence of AP and DS from the model indicates that omitted variable bias can be expected whenever these ability factors are studied in isolation from SS and AC. The practical implication is that physical training can be mistakenly focused on abilities that have no impact on job performance.

15. SUBJECT TERMS
physical ability, job performance, strength, aerobic capacity, lifting, carrying, omitted variable bias

16. SECURITY CLASSIFICATION OF:

a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
UNCL	UNCL	UNCL	UNCL	26

18a. NAME OF RESPONSIBLE PERSON
Commanding Officer

18b. TELEPHONE NUMBER (INCLUDING AREA CODE)
COMM/DSN: (619) 553-8429